A Learner's Guide to the Semantic Web

Briefing to CENDI - 11/4/2010

George O. Strawn
Director of the NCO
Co-chair of NITRD

Outline

- what is: the web? semantics? the semantic web?
- what's new about the semantic web?
- global names, rdf triples and triple stores
- tables and text as rdf triples
- graphical representation of rdf triples
- semantic search with sparql
- inferred triples via rdfs and owl
- bottlenecks and breakthroughs
- conclusions

What is the web?

- web pages *linked together*

- a web page has *information for humans* to read with hidden metadata (html) to make the it more readable

- web search is performed by specifying key words (and then retrieving a million pages)

- by contrast, the semantic web has *data for computers* to read and semantic searches yield answers, not pages to read that may have answers

Example web page

http://www.nsf.gov/grant

(grant number principle investigator dollar amount in Ks)

GRANT	PI	AMT
1	smith	100
2	jones	100
3	miller	200

a google search for "smith" yield a pointer to the whole page

a *sparql* query for (?g nsf:PI smith.) yields ?g = 1;

What is semantics?

- syntax refers to form; semantics refers to meaning

- but what does *meaning* mean?

- stay tuned to see what meaning means for the semantic web

- viewpoint: I might have called it “the inferred web” or “the computed web” or “the atomic web”

What is the Semantic Web?

- Is it the latest new, new thing?
- Is it the greatest IT invention of all time?
- A system that will solve all your IT problems?
- No, but it *is* a remarkable new way to *federate data* (combine data sets, merge data, do mashups, etc)

  Since I never understand “what it can do” without understanding “how it does it”, you will get some of both from me today

What's new about the semantic web?

- "The only thing new about the semantic web is the web” (ie, computer scientists have been studying semantics for decades)

- The semantic web has "meaningful data", which means that computers can "understand" it better than plain web data (ie, do more with it)

nb - Much as the web links pages to pages, the semantic web *links data elements to data elements* (“nouns linked to nouns by links labeled by verbs”)

global names and rdf triple stores

- Web URLs become URIs to create *globally unique names* for identified nouns and verbs in a text (or table)

- These uniquely named nouns and verbs are the parts of the “key sentences” of the text (also of table elements) and are of the form *subject predicate object*

- The semantic web calls these forms *rdf triples*

- A semantic web database is a set of rdf triples, called a *triple store*

a namespace example for a table

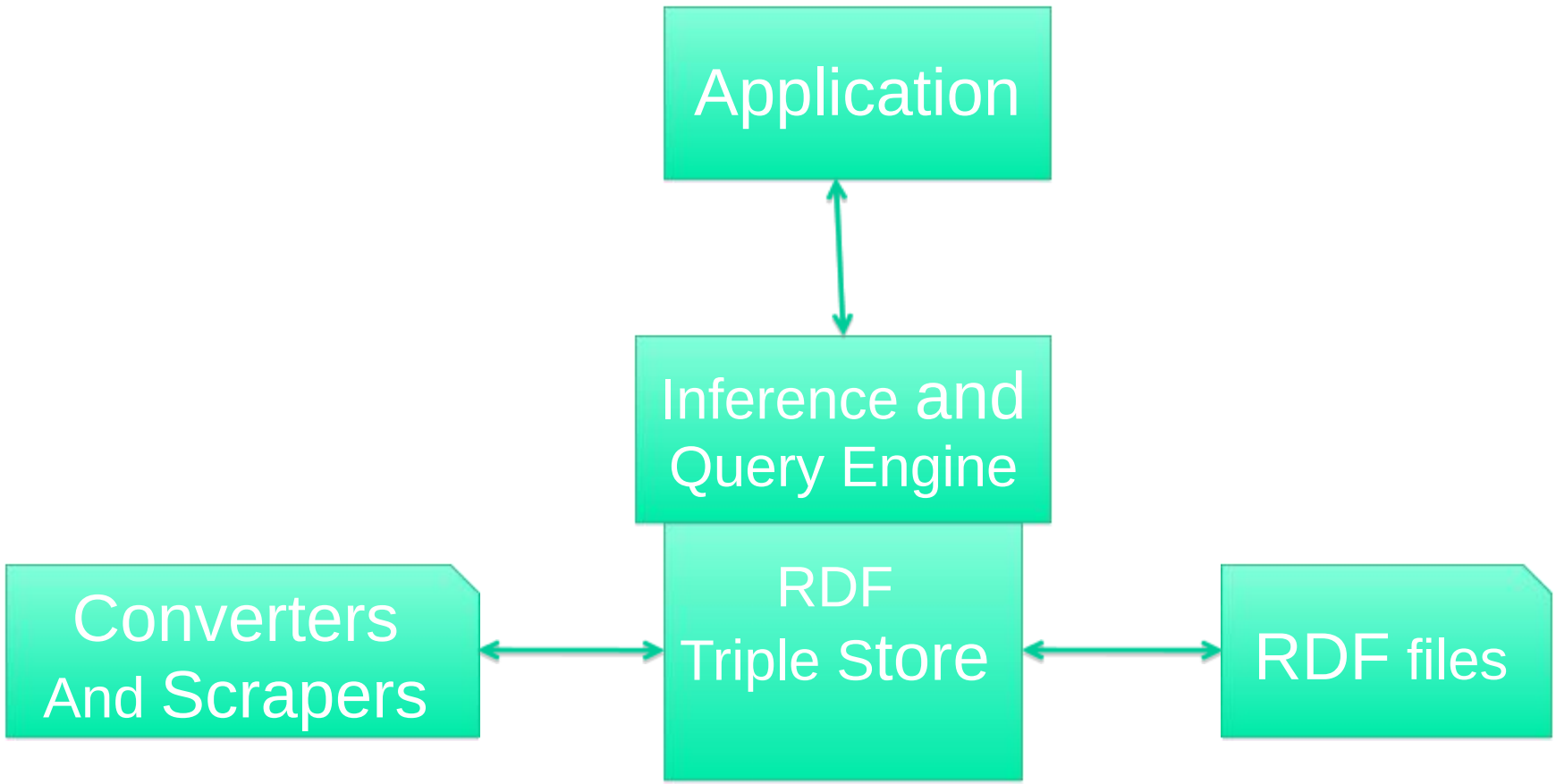
http://www.nsf.gov/grant

GRANT	PI	AMT
1	smith	100
2	jones	100
3	miller	200

---

if < rdf xmlns: nsf = http://www.nsf.gov/grant# >
then http://www.nsf.gov/grant#PI becomes nsf:PI
and http://www.nsf.gov/grant#AMT becomes nsf:AMT

The Semantic Web System



Text and tables as rdf triples

- Text: the rdf triples are *key sentences* (of the form subject verb object), which are to be extracted from the text by natural language processing

- Keyed tables: the rdf triples are of the form:
key name_key value (subject),
column name (predicate),
table value for that key and column (object)

- note that the rdf for a text is smaller than the text and the rdf for a table is larger than the table

- nb: both text and table become sets of rdf triples

rdf triples from text example

NSF competitively awards grants in all branches of science and engineering. NSF challenges its awardees to be at the cutting edge of their disciplines. Ideally, the projects will lead to transformative research.

NSF *makes* grants

NSF awardees *perform at* the cutting edge

NSF projects *should be* transformative research

rdf triples from a table

GRANT (key)	PI	AMT
1	smith	100
2	jones	100
3	miller	200

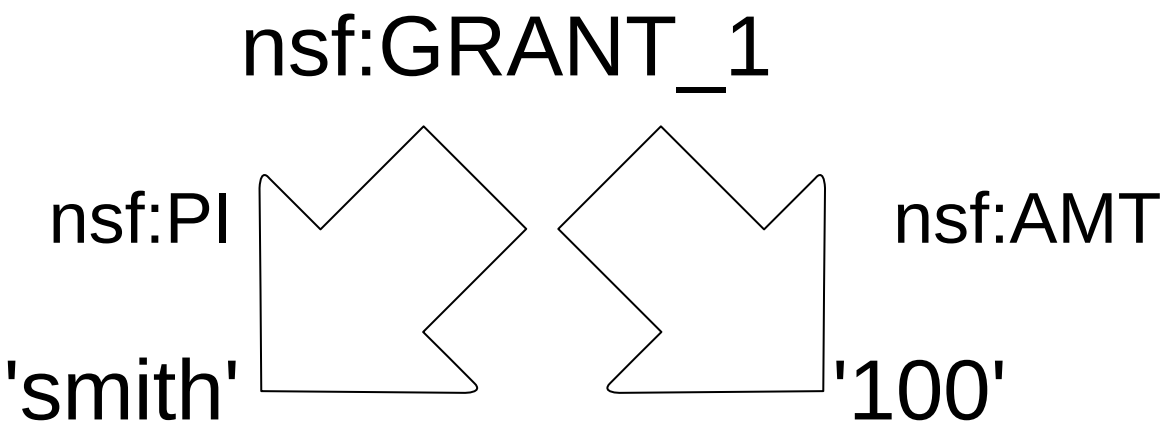
nsf:GRANT_1 nsf:PI smith
nsf:GRANT_1 nsf:AMT 100

nsf:GRANT_2 nsf:PI jones
nsf:GRANT_2 nsf:AMT 100

nsf:GRANT_3 nsf:PI miller
nsf:GRANT_3 nsf:AMT 200

Graphical representation of rdf triples

- create a graph whose nodes are labeled by subjects and objects of the key sentences and the arcs are labeled by the predicates



Semantic search with sparql

-key sentences beat key words for searching!

- search for triples with variables or for combinations of triples
- sparql query (?g nsf:PI 'smith'.)

?g = nsf:GRANT_1;

- sparql query (?g nsf:AMT ?a. ?h nsf:AMT ?a.)

?g=nsf:GRANT_1; ?h=nsf:GRANT_2; ?a=100

Inferred triples via rdfs and owl

- rdfs and owl are “instruction” triples, the “execution” of which produces more triples (inferred triples)

- thus, an rdf triple store can have asserted triples, instruction triples and inferred triples

and…

-*classes* are *sets* of rdf subjects/objects and *properties* are *sets* of rdf predicates.

- inferencing is as much about classes and properties as it is about rdf triples

inferred triples example

1. (?g PI ?x.) -> ?g=grant_1; ?x = smith
   ?g=grant_2; ?x = jones
   ?g=grant_3; ?x = miller

2. 'smith' *type* researcher
   'jones' *type* researcher -> (?x type researcher)
   'miller' *type* researcher
   ?x=smith;
3. PI *range* researcher ?x=jones;
   ?x=miller;

Inference Rule to infer #2:

If x PI y then y *type* researcher

An Ontology: classes and predicates

classes are sets of nouns (individuals)
predicates are sets of verbs (properties)

researcher was an example of a class
smith et al were the individuals in that class

PI was an example of a property

An ontology can be thought of as a graph whose nodes are classes and whose arcs are labeled by properties

eg, GRANTS --PI--> researchers

Bottlenecks and breakthroughs

- bottleneck: too big a change of concept to "satisfy current user needs”

- bottleneck: vocabulary infrastructure (exception: semantic medline)

- breakthrough: "linked data" (tables only) (eg, DBpedia)

- breakthrough: linguistic progress (eg, semantic medline)

Conclusions

- The semantic web offers the possibility of a unique way to federate datasets of widely differing formats

- The semantic web may be ready to move from the experimental phase to the early adopter phase

- But as Yogi's said:

Predictions are hard--especially about the future